

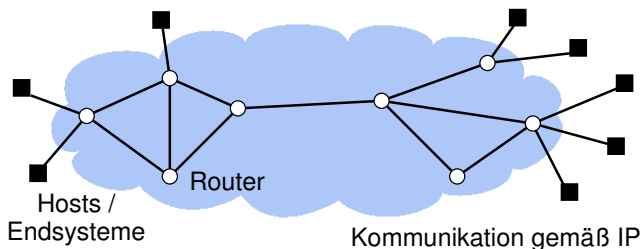
**Einführungsvortrag:  
Webgraph, Klassisches IR vs. Web-IR  
Seminar Suchmaschinen,  
Wintersemester 2007/2008**

**Martin Sauerhoff**

**Lehrstuhl 2, Universität Dortmund**

1. Einleitung
2. Der Webgraph
3. Modelle für den Webgraphen
4. Klassisches IR vs. Web-IR

## Das Internet – abstrakte Version:

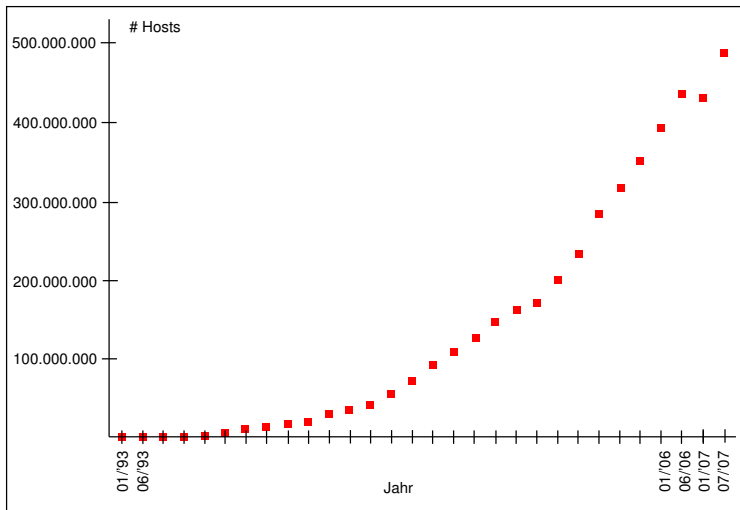


In Wirklichkeit starke hierarchische Struktur,  
viele verbundene Einzelnetze.

- Internationale Verbindungen;
- nationale Backbones;
- ISP-Backbones;
- lokale Netze.

# Wie groß ist das Internet?

Aktuelle Schätzung für Anzahl Hosts: ca. 490 Millionen  
(gemäß *Internet Systems Consortium*, [www.isc.org](http://www.isc.org)).



## Wie groß ist das Internet? (Forts.)

Wie das ISC zu diesen Zahlen kommt:

- „Hosts“: Geräte mit global eindeutiger IP-Nummer.
- Zähle alle IP-Nummern, zu denen es einen Host gibt.  
Dazu Abbildung IP-Nummer → Hostname per  
inversem DNS-Lookup.
- Insgesamt  $2^{32}$  IP-Nummern, daher vorab „Pruning“  
von nicht erreichbaren Teilnetzen.

Details: `www.isc.org/ops/ds/rfc1296.txt`.

## WWW (aka: Web, Netz, Internet (?)):

- Realisiert mittels HTTP (Anwendungsschicht).
- Virtuelles Netz aus HTML-Dokumenten (Webseiten) mit Hyperlinks als Verbindungen.

### Definition 2.1: Webgraph

Gerichteter Graph  $G = (V, E)$  mit

- Knotenmenge  $V$ : statische Webseiten.
- Kantenmenge  $E$ : Hyperlinks.

Dabei nicht erfasst: Dynamische Webseiten.

- Generierung bei Anfrage durch Skript (PHP...).
- Oft: Zugriff auf Datenbanken.

# Suchmaschinen:

Benutzeranfrage  $\mapsto$  Liste von Referenzen auf Webseiten.

- *Crawler*: Regelmäßiger, automatischer Durchlauf (*Crawl*) des Webs zwecks Datensammlung. Liefert:
- *Index*: Abbildung Webseiten-ID  $\mapsto$  Inhalt.
- **Surface Web, indizierbares Web**: Webseiten, die prinzipiell für Crawler von Suchmaschinen erreichbar sind. Nur ein kleiner Teil davon tatsächlich indiziert.
- **Deep Web**: Der Rest.
  - Dynamische Seiten, insbesondere nichtöffentliche Datenbanken.
  - Seiten, die Crawler ausschließen (`robots.txt`).
  - Isolierte (nicht verlinkte) Seiten.

## Wie groß ist das Web?

Alles Schätzungen, oft ideologisch/ökonomisch motiviert.

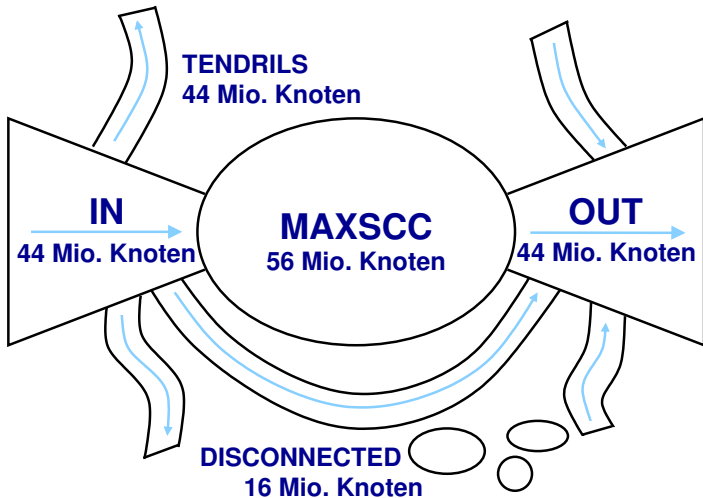
- **Anzahl Web-Server:** 143 Mio. (Oktober 2007)  
Siehe `news.netcraft.com`.
- **Indiziertes Web:** 10 – 30 Mrd. Seiten.
  - Letzte offizielle Angabe von Google (2005): 8 Mrd.
  - Suchbegriff „a“ bei Google: 15 Mrd. ;–)
  - Schätzung aus Web-Server-Anzahl: 29,7 Mrd.  
(`www.boutell.com`, Februar 2007)
- **Indizierbares Web:**  
Gulli, Signorini (2005): 11,5 Mrd. Seiten  
Schätzung mit Sampling-Techniken.

### Eigenschaften des Webgraphen (aus Experimenten):

- „Fliegenstruktur“ des Webgraphen;
- „Kleine-Welt-Phänomen“;
- Webgemeinden;
- Potenzgesetze z. B. für Ein- und Ausgangsgrad.

## Fliegenstruktur des Webs (Broder u. a. 2000):

Analyse eines Altavista-Crawls von 1999 mit  
204 Mio. Webseiten, 1,5 Mrd. Links.



## Kleine-Welt-Phänomen

Ursprünglich für soziale Netze (Milgram-Experiment, 1966).  
Mathematische Formalisierung über Durchmesser.

### Definition 3.1: Durchmesser

Betrachte gerichteten Graphen  $G = (V, E)$ .

- *Durchmesser,  $\text{diam}(G)$* : Maximale Länge eines kürzesten Weges zwischen zwei Knoten.
- *Durchschnittlicher Durchmesser,  $\overline{\text{diam}}(G)$* : Durchschnitt der kürzesten Weglänge über alle **verbundenen** Knotenpaare.

Falls  $\overline{\text{diam}}(G)$  endlich, dann höchstens  $n - 1$ ,  $n := |V|$ .

Populäre Formalisierung des Kleine-Welt-Phänomens:

$$\overline{\text{diam}}(G) = O(\log n).$$

# Durchmesser des Webgraphen

## Arbeit von Broder u. a. (2000):

Altavista-Crawl von 1999:

- $\text{diam}(\text{MAXSCC}) \geq 28$ .
- Maximale endliche Weglänge vermutlich ca. 900.
- Nur für 24 % aller Knotenpaare  $(v, w)$  existiert Weg  $v \rightsquigarrow w$ . Ausschluss von nicht verbundenen Paaren bei durchschnittlichem Durchmesser wichtig.
- BFS-Durchläufe für Startknoten-Sample liefern:

	Vorwärtskanten:	Rückwärtskanten:	Ungerichtet:
diam:	16,12	16,18	6,83

## Webgemeinden

### Webgemeinde:

Gruppe von im Web repräsentierten Individuen mit gemeinsamen Interessen zusammen mit gemeinsam von ihnen bevorzugten Webseiten.

Etablierte Webgemeinden über Newsgroups, Verzeichnisse wie Yahoo!, `web.de` usw.

Ziel: Automatisches Erkennen von „aufstrebenden“ Webgemeinden anhand von Linkstruktur.

**Arbeit:** Kumar, Raghavan, Rajagopalan, Tomkins (1999).

## Wie soll das gehen?

### **Autoritäten (*authorities*):**

Thematisch verwandte Seiten, für die sich Webgemeinde interessiert.

**Problem:** Oft keine direkten Links zwischen solchen Seiten!

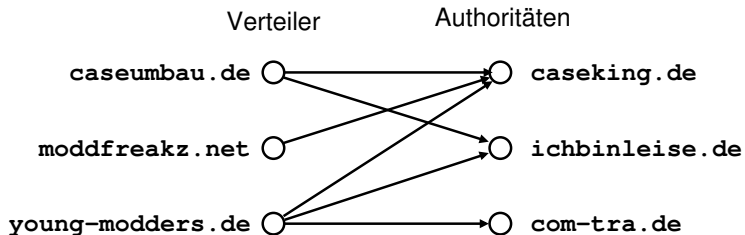
**Beispiel:** Webshops für Case-Modding-Zubehör.

**Idee:** Identifiziere Seiten, die Links auf (viele) autoritative Seiten haben.

### **Verteiler (*hubs*):**

Seiten, die Links auf Autoritäten der Webgemeinde haben.

## Winziger Ausschnitt Case-Modder-Gemeinde:



## Formalisierung:

Verteiler und Autoritäten einer Webgemeinde:

- Bipartiter Teilgraph des Webgraphen.
- Ignoriere evtl. vorhandene (wenige) Links innerhalb der Verteiler / Autoritäten.

Vollständiger bipartiten Teilgraph:

*Kern (core)* der Webgemeinde.

### Definition 3.2: Vollständiger bipartiter Graph

Graph  $K_{i,j}$ :

- Knotenmenge  $V_1 \dot{\cup} V_2$  mit  $|V_1| = i$ ,  $|V_2| = j$ ;
- Kantenmenge  $E = V_1 \times V_2$ .

## Experimente von Kumar u. a. (1999)

**Daten:** Alexa-Crawl von 1998, ca. 200 Mio. Seiten.

### Ergebnisse:

- In 1999: Newsgroups, kommerzielle Verzeichnisse: ca. 10.000 etablierte Gemeinden.
- Arbeit liefert ca. 100.000 aufstrebende Gemeinden.

### Kern-Anzahlen (Ausschnitt):

$(i, j):$	$(3, 3):$	$(4, 3):$	$(5, 3):$	$(6, 3):$
$\# K_{i,j}:$	38.887	11.410	7.015	2.757

### Erkenntnis über Webgraphen:

Existenz von großer Anzahl von  $K_{i,j}$  für kleine  $i, j$  (Clusterbildung).

# Potenzgesetze

## Beobachtung:

Viele Parameter in natürlichen / sozialen Systemen zeigen Streuung über großen Wertebereich.

**Insbesondere:** Gerade *nicht* normalverteilt.

## Beispiele:

- Einkommensverteilung;
- Einwohnerzahlen;
- Worthäufigkeiten;
- Anzahl eingehender / ausgehender Links auf Webseiten.

Verteilungen folgen *Potenzgesetzen*.

### Definition 3.3: (Diskrete) Potenzgesetz-Verteilung

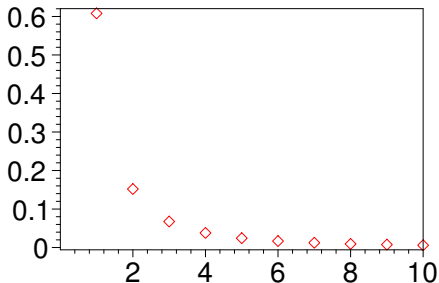
Sei  $\alpha > 1$ . Zufallsvariable  $X \in \mathbb{N}$  hat (*diskrete*) *Potenzgesetz-Verteilung*, falls

$$\Pr\{X = x\} = \frac{c}{x^\alpha}, \quad \text{wobei} \quad \frac{1}{c} = \sum_{x=1}^{\infty} \frac{1}{x^\alpha} = \zeta(\alpha).$$

Nenne  $\alpha$  *Exponent* der Verteilung.

Alternative Namen: *Zipf-Verteilung*, *Zeta-Verteilung*.

**Beispiel:** Exponent  $\alpha = 2$ .



$$\frac{1}{c} = \sum_{x=1}^{\infty} \frac{1}{x^2} = \frac{\pi^2}{6}.$$

## Log-Log-Plots

Sei  $f(x) := c/x^\alpha$ . Dann:

$$\log_{10} f(x) = \log_{10} c - \alpha \cdot \log_{10} x;$$

$$\log_{10} f(10^{\tilde{x}}) = \log_{10} c - \alpha \cdot \tilde{x}.$$

Allgemein: Plot von  $\tilde{x} \mapsto \log_{10} f(10^{\tilde{x}})$  heißt *Log-Log-Plot*.

Liefert für  $f =$  Potenzgesetz-Verteilung *lineare Funktion* mit Steigung  $-\alpha$  und vertikaler Verschiebung  $\log_{10} c$ .

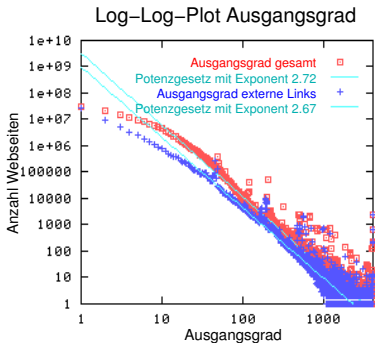
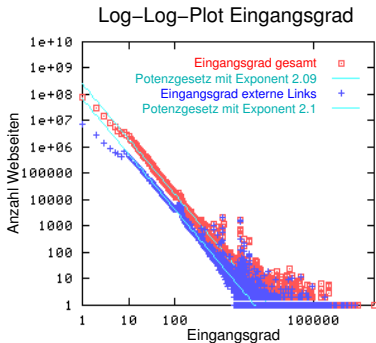
In Anwendungen oft einfach Plot der *absoluten Häufigkeiten*, für Parameter  $c$  dann beliebige Werte zulässig.

# Webgraph: Potenzgesetze für den Knotengrad

## Arbeit von Broder u. a. (2000):

Untersuchung der Eingangsgrade und Ausgangsgrade, d. h. Anzahl eingehender bzw. ausgehender Links. Daten:

- Altavista-Crawl von 1999.
- Zusätzlich externe Seiten mit mindestens 5 Links aus ursprünglichem Crawl.



## Weitere Potenzgesetze im Internet:

- Größen der schwachen und starken ZKs des Webgraphen (Broder u. a. 2000; Donato u. a. 2004);
- Anzahl Benutzerzugriffe auf verschiedene Web-Sites (Adamic und Huberman 2001);
- Gradverteilung für Internet-Graph auf verschiedenen Hierarchieebenen (Faloutsos u.a. (1999) + Nachfolgearbeiten, z. B. Chen u. a. (2002), Jaiswal u. a. (2004)).

## 3. Modelle für den Webgraphen

### Wozu überhaupt?

- Test von Algorithmen;
- Vorhersage zukünftiger Entwicklung;
- besseres Verständnis beobachteter Phänomene.

### Anforderungen an Modelle:

- Dynamische Entwicklung;
- Potenzgesetze, z. B. für Eingangsgrad;
- kleiner durchschnittlicher Durchmesser;
- Clusterbildung, insbesondere viele  $K_{i,j}$ -Kopien.

## Einige wichtige Modelle:

- **ER-Modell (Erdős, Rényi 1960):**

Kanten zufällig unabhängig mit vorgegebener Wskt.  
Knotengrade, Clusterbildung nicht richtig, keine Dynamik.

- **Preferential Attachment (Barabási-Albert 1999):**

Wskt. für Verbindung von neuem Knoten zu altem Knoten proportional zu dessen Grad. Erzeugt nur Bäume.  
Aber: Potenzgesetz für Gradverteilung.  
Viele Erweiterungen (insbes. [Aiello, Chung, Lu \(2001\)](#)).

- **Kopiermodelle (Kumar u. a. 2000):**

Einfache Variante: Neuer Knoten kopiert viele Kanten von altem Knoten und erzeugt einige Kanten zufällig neu.  
Potenzgesetze, Existenz großer bipartiter Cliques.

## 4. Klassisches IR vs. Web-IR

Zunächst einige (wenige) Grundbegriffe zu klassischem Information-Retrieval.

Wesentliche Schritte beim IR:

- Datenaufbereitung  $\rightarrow$  Index / inverser Index.
  - **Index:**  
Abbildung Dokumente  $\mapsto$  enthaltene Terme  
(Dokumente  $\leftrightarrow$  Webseiten, Terme  $\leftrightarrow$  Suchbegriffe);
  - **inverser Index:**  
Terme  $\mapsto$  Dokumente, die diese enthalten.
- Anfrageauswertung.
- Ranking der Ergebnisse.

Viele Techniken zur Verarbeitung von natürlichsprachlichen Texten, auch für Suchmaschinen relevant.

## Anfrageauswertung:

Typische Verfahren aus der Computerlinguistik:

- Zerlegung von Anfragen in Einzelwörter;
- Entfernen von Stoppwörtern (Artikel, Konjunktionen. . . );
- *Stemming*: Reduktion von Suchbegriffen auf Grundform.  
Beispiel: Spam-mer, Spam-ming → Spam.

## Ranking:

Suchergebnisse absteigend sortieren nach Relevanz zur Anfrage.

# Inhaltsbasiertes Ranking

## Vektormodell für Dokumente: 5.1:

- Vokabular  $V$ ,  $|V| = n$ . Elemente: Mögliche Suchbegriffe.
- Dokument  $d \mapsto$  Vektor  $w(d) = (w(d)_1, \dots, w(d)_n) \in \mathbb{R}^n$ ;  
 $w(d)_i$ : Gewicht des  $i$ -ten Begriffes im Dokument  $d$   
(z. B.  $w(d)_i = \#$  Vorkommen von Begriff  $i$  in  $d$ ).
- *Kosinusmaß für Ähnlichkeit:*

$$\begin{aligned}\text{sim}(d_1, d_2) &= \frac{\langle w(d_1), w(d_2) \rangle}{\|w(d_1)\|_2 \cdot \|w(d_2)\|_2} \\ &= \cos(\angle(w(d_1), w(d_2))) \in [-1, 1].\end{aligned}$$

## Beispiel:

$$w(d_1) = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, w(d_2) = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \Rightarrow \text{sim}(d_1, d_2) = \frac{2}{\sqrt{6}} \approx 0.82.$$

## Inhaltsbasiertes Ranking (Forts.)

Betrachte Gesamtmenge von Dokumenten

$D = \{d_1, \dots, d_k\}$  (z. B. Web). Sei  $i \in \{1, \dots, n\}$ ,  $n = |V|$ .

- **Term-Frequency, TF:**

$TF(d)_i := \#$  Vorkommen des  $i$ -ten Begriffes in  $d$ ;

- **Document-Frequency, DF:**

$DF_i := \#$  Dokumente, die  $i$ -ten Begriff enthalten;

- **Inverse-Document-Frequency, IDF:**

$IDF_i := \log(k/DF_i)$  (setze  $DF_i \neq 0$  voraus).

**TF-IDF-Gewichte:**

$$w(d)_i := TF(d)_i \cdot IDF_i, i = 1, \dots, n.$$

## Inhaltsbasiertes Ranking (Forts.)

### Beispiel:

Begriff:	$TF(d_1)_i:$	$TF(d_2)_i:$	$TF(d_3)_i:$	$IDF_i:$
GraKa	12	8	1	$\log(3/3) = 0$
SuSE	2	0	1	$\log(3/2) \approx 0.58$
funzt	0	1	0	$\log(3/1) \approx 1.58$

$$TF-IDF(d_1) = (12 \cdot 0, 2 \cdot 0.58, 0 \cdot 1.58) = (0.00, 1.16, 0.00);$$

$$TF-IDF(d_2) = (8 \cdot 0, 0 \cdot 0.58, 1 \cdot 1.58) = (0.00, 0.00, 1.58);$$

$$TF-IDF(d_3) = (1 \cdot 0, 1 \cdot 0.58, 0 \cdot 1.58) = (0.00, 0.58, 0.00).$$

## Web-Information-Retrieval: Was ist anders?

Unterschiede Web ↔ klassische Datenbanken:

- Datenmenge (siehe Einleitung).
- Dynamik (z. B. 11 Tage, bis 50 % der Seiten in .com in Web-Crawl geändert ([Cho, Garcia-Molina 2000](#))).
- Wenig Fließtext, dafür viele andere Medien.
- Vielzahl von Sprachen.
- Spam.
- Hoher Gehalt an Verbindungsinformation (Hypertext).

## Ziele für Suchmaschinen?

Klassische IR-Begriffe:

- Genauigkeit (Precision):

Anteil gefundener & relevanter Dokumente an allen *gefundenen* Dokumenten.

- Vollständigkeit (Recall):

Anteil gefundener & relevanter Dokumente an allen *relevanten* Dokumenten.

Für Web:

- Hohe Vollständigkeit illusorisch und auch nicht erwünscht.
- Vielzahl an relevanten Dokumenten.

## Ziele für Suchmaschinen (Forts.):

Beobachtung Benutzerverhalten:

- Unpräzise Benutzeranfragen (nur 1-3 Suchbegriffe).
- Nur selten mehr als erste Seite Ergebnisse betrachtet.

Ziele für Web-IR (Henzinger 1999, damals bei Google):  
Relevanz & Qualität (in den Augen des Benutzers),  
qualitativ hochwertige Ergebnis unter ersten Ergebnissen.

Im Vergleich zu klassischem IR neue Ideen  
für Ranking erforderlich.

## Problem für Suchmaschinen – Web Spam:

- Bewusste Herbeiführung von Überbewertung des Seiteninhalts durch Suchmaschinen.
- Schwierig: Abgrenzung gegen legitime Suchmaschinen-Optimierung (SEO).

### **Schätzungen für 2004 (lt. Gyöngyi, Garcia-Molina 2005):**

- 15–18 % Spam in Suchmaschinen-Indexen;
- 9 % Suchergebnisse mit Spam in Top-10.

Arbeit liefert auch Klassifikation von Spam-Techniken.